Computational Future of in silico Science

Lessons from the Genomic Revolution

Biology, Not Physics, Now Drives Computing Technology

"By the year 2020, petacrunchers, capable of reaching a thousand trillion calculations per second, may be possible, although new technologies and programming methods will be needed to reach that level."

"At this point, brute-force simulation of cell mechanics, tracking every active molecule and its web of interactions, should be attainable."

- Prof. Edward O. Wilson, Harvard University

What Celera Anticipated

- Assembly
  - 1 Terabyte memory
  - 6-8 Alpha processors
  - 6 – 8 Months
- GCMP
  - Discrete Systems
  - 100 processors
- Target → September 2002
- 90,000 to 120,000 genes
- 4-6B base pairs
- Data Ownership

What Transpired

- Assembly
  - 32 GB
  - Two Assemblers (two separate approaches)
  - New Assembler per mammalian genome
  - 100+ loosely coupled CPUs
  - 20,000 CPU hours
  - 20+ Human Assemblies
- GCMP
  - Shared Systems
  - 800 processors
  - 100+ TB
- Target → May 2000
- 30,000 genes
- 3.12B base pairs
- Data Explosion - Annotation

The Bottom Line

Biology, specifically the nature of disease, is much, much more complex than originally anticipated. It is as much a data management and access challenge as it is a compute challenge.

- No one can own all the data needed
- The data is distributed
- No one yet knows what data is needed
- Policy Challenge

What is Needed

A flexible, scalable, secure, highly performing, highly available computing infrastructure that adapts to a wide range of continuously evolving and challenging demands.

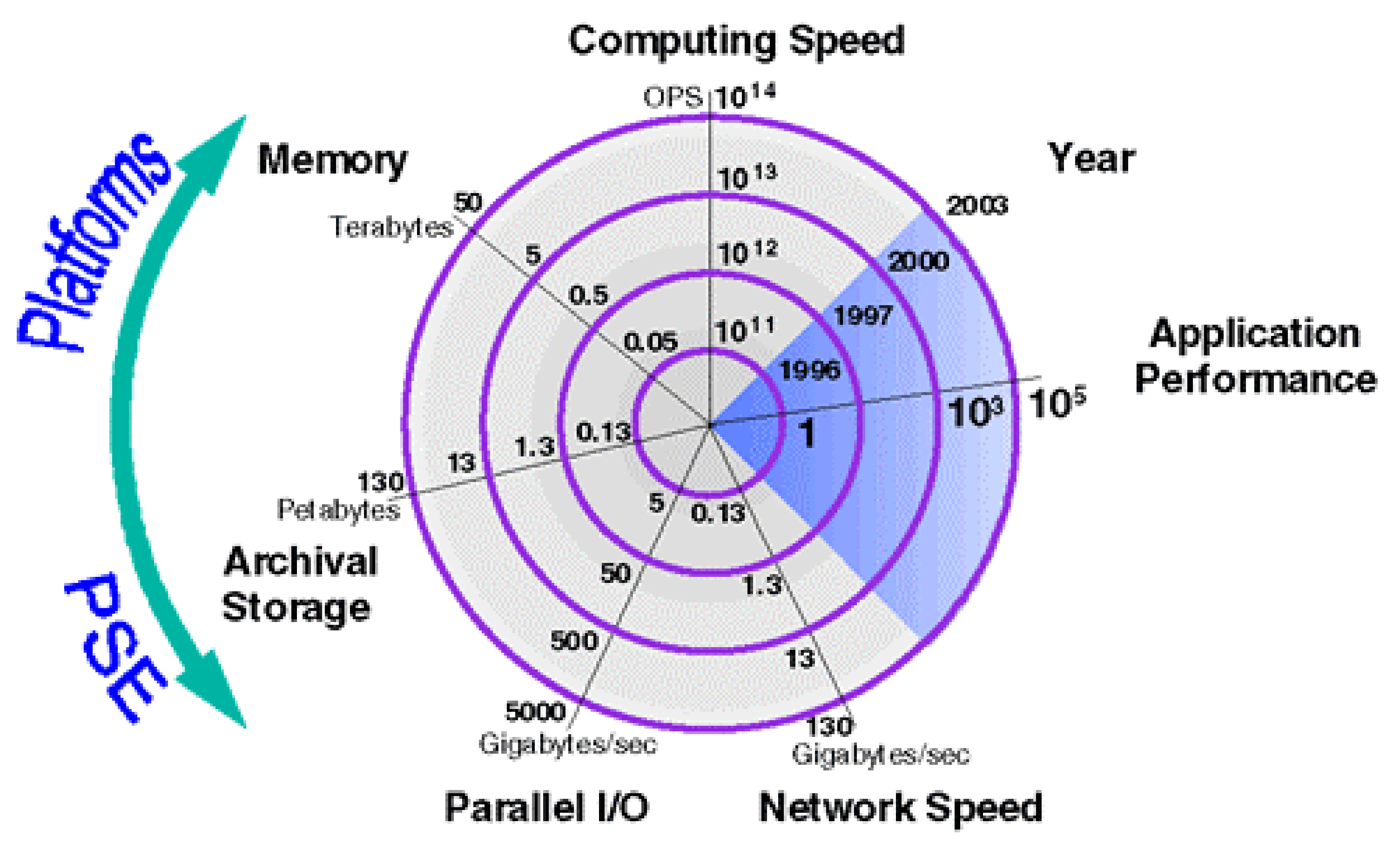
Infrastructure Challenges

- Flexibility
- Implementation Speed
- Extremely High File System Throughput
- Ease of Use
- Scalability of Capacity
- Technology Refresh
- People

In vivo, in vitro, and now in silico…

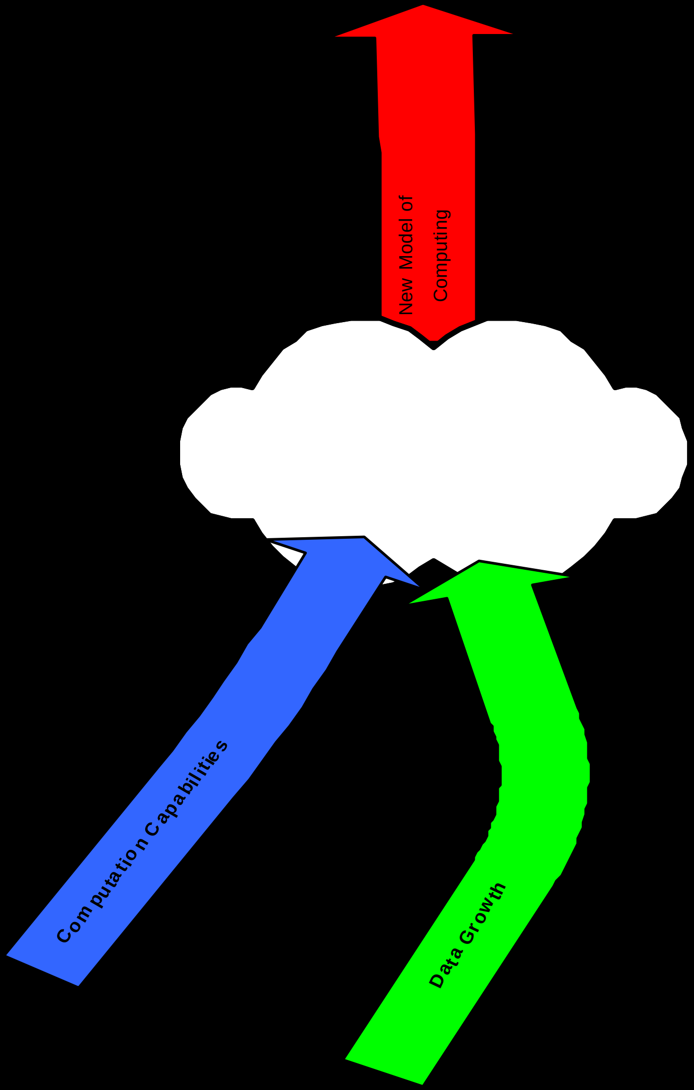
- An unprecedented combination of supercomputing and database search techniques were deployed to sequence and assemble the human genome, yet orders of magnitude improvement are needed.
    - Highly parallel supercomputers and database systems are becoming “commodity components” in the overall system while evolving and adapting in key ways
    - IT innovation will intensify at the system architecture level

DOE ASCI



The U.S. Dept. of Energy Accelerated Strategic Computing Initiative
Drove Application Scaling X10000 in the Past 6 Years

Data & Computing



New Model of Computing

Computation Capabilities

Data Growth

Data Driven Computing

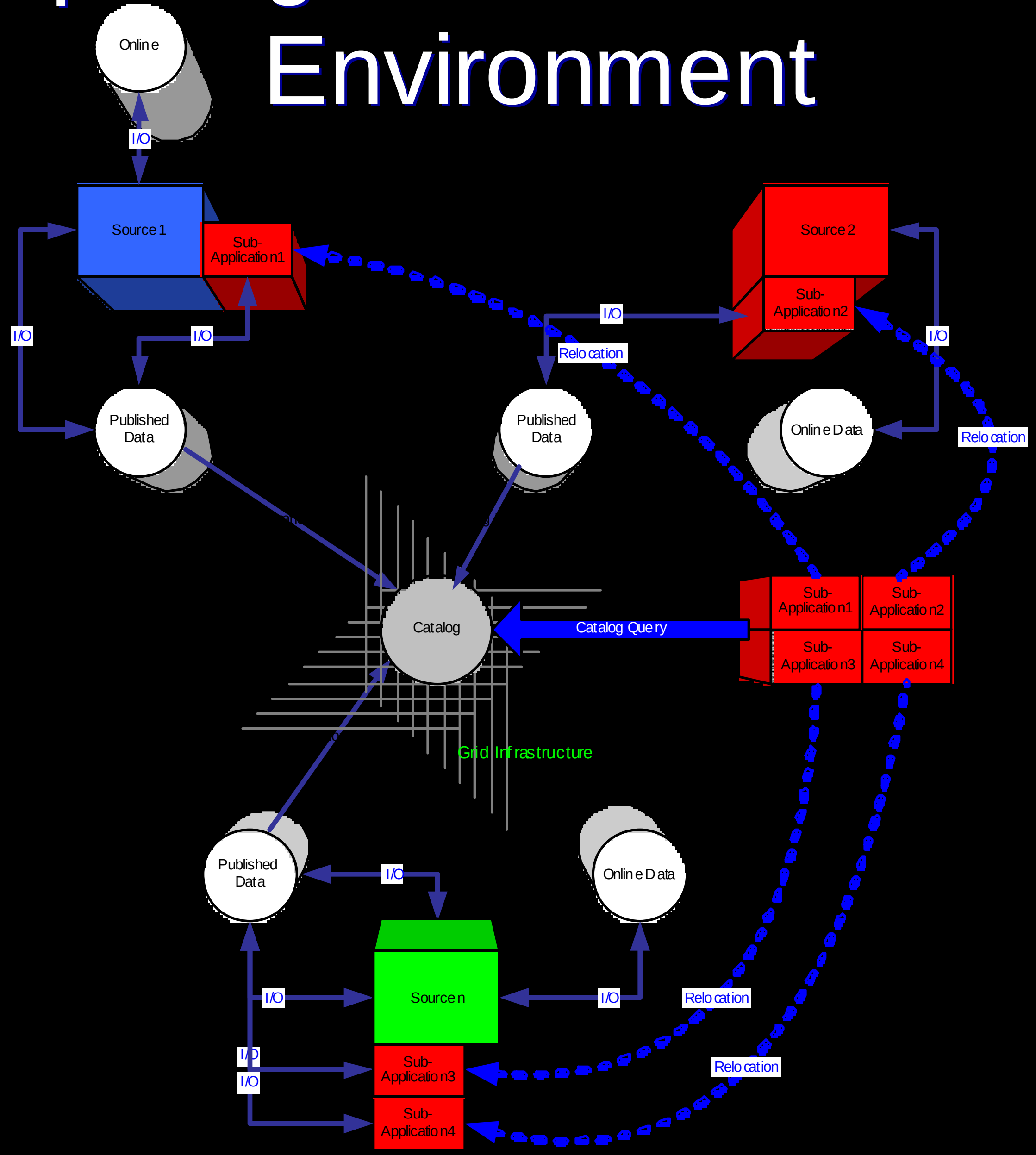
We are at the beginning of the genomic era where the linking of genomic, proteomic, environmental, clinical, and a myriad of other information will revolutionize our understanding of ourselves and our environment . Systems capable of efficiently processing massive amounts of heterogeneous data from a variety of sources will form the foundation of this revolution. These data will cover a large range of scales in time, location, and space, from the molecular level through to individual patients, to populations and the environment. This will require new concepts in large scale distributed data management. The development of these systems is best undertaken through collaboration between vested parties.

Data Centric vs. Computation

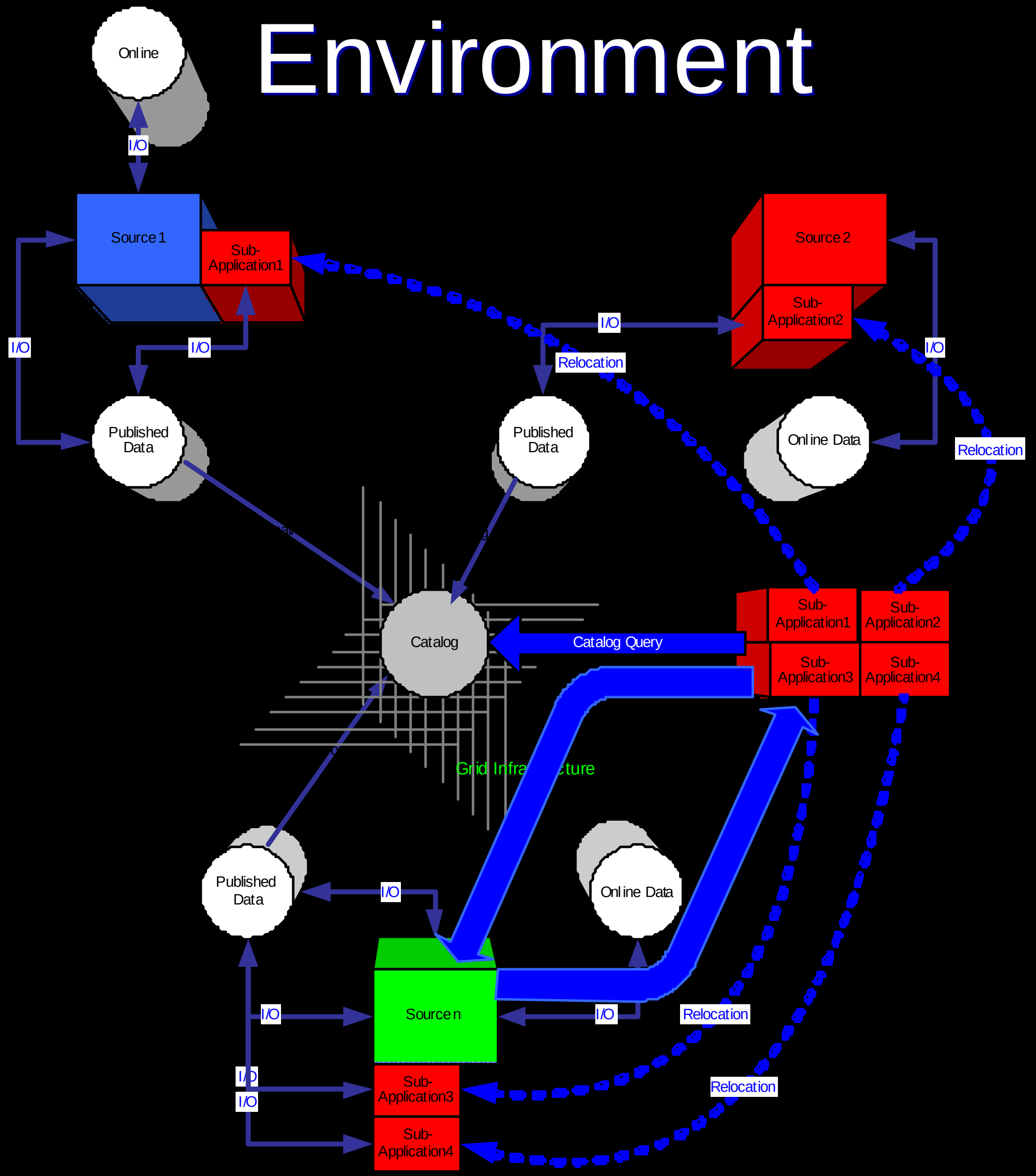
The primary requirement is the management of massive amounts of heterogeneous data in a very distributed environment. By management we mean capture, storage, archival, distribution, cataloging, retrieval, and analysis.

Current and planned designs for highly-scalable computing favor the scheduling and distribution of computation as opposed to the scheduling and distribution of data. The next-generation of these pipelines and the software and hardware systems that support them must be structured to minimize the penalty of moving the data to the computational resource. In practice this means that computation must balance data location with the availability of appropriate compute resources.

Computing in a Distributed Data Environment



Computing in a Distributed Data Environment


Online
I/O
Source 1
Sub-Application1
I/O
I/O
Source 2
Sub-Application2
I/O
Relocation
I/O
Published Data
Published Data
Online Data
Relocation
Catalog
Catalog Query
Sub-Application1
Sub-Application2
Sub-Application3
Sub-Application4
Grid Infra cture
Published Data
I/O
Online Data
I/O
Source n
I/O
Relocation
I/O
I/O
Sub-Application3
Sub-Application4
Relocation